最適停止をともなう決定過程 (制御過程論 II)

AUTHOR(S):

古川, 長太

CITATION:

古川, 長太. 最適停止をともなう決定過程 (制御過程論 II). 数理解析研究所講究録 1970, 90: 171-207

ISSUE DATE:

1970-06

URL:

http://hdl.handle.net/2433/108115

数理解析研究所講究録
90 巻 1970 年 171-207

Ⅳ 決定過程

最適停止をともなう決定過程

九大 理 古川長太

§1. 序

離散時間確率的系に対する最適化問題では，stage 数を予め指定せずに無限 stage にわたっての政策を問題とするならば，通常，評価関数として無限和（stage についての和），またはその 1 stage 当りの平均の型を採用する。ここでは無限和の型の評価関数を最大化することに限って考えることにする。

無限和の評価関数を採用することは，一応，我々が無限 stage にわたって行動し続けることを前提としているが，現実にはいつかは stop せざるを得ない。また，むしろ stop した方がかえって有利な場合もある。また，いわゆる sequential analysis のように terminal decision が要求されるときは，当然いつかは stop しなくては問題の

解決にならない。

ここに一つの決定過程($\mathbb{P}$)が与えられたとし，$\mathbb{P}$の状態空間をS，行動空間をAとする。Sに仮想的な状態s_0(吸収点)を加え，Aにstopを表わす行動a_0を加えて，$S' = S + \{s_0\}$，$A' = A + \{a_0\}$を新たにそれぞれ状態空間，行動空間として新しい決定過程($\mathbb{P}'$)を構成すると，$\mathbb{P}'$はstopするという行動をも含んだ決定過程になる。$\mathbb{P}'$に対して無限和の評価関数を採用すると，表面上は無限和であるがa_0をとった時点でprocessは実際にはstopし，したがって，はじめに述べた問題は，このような定式化により通常の最適化問題に帰着されるかのように見える。しかしながら，$\mathbb{P}'$における最適政策が$\mathbb{P}$において果たして確率1でstopすることになるかどうか分らないし，また$\mathbb{P}$でstopする確率を計算することも一般に仲々困難である。

以上の理由により，ここに全く新しい定式化で考えなおすことにする。この報告では"確率1でstopする政策"(非常にroughな表現であるが)を考え，そのような政策の集合の中で最適なものを追究する。正確な定義は次の章にゆずるとして，さらに詳しく述べれば，通常の意味の政策πと，停止時間τをpairとして考え，あらゆるpair (π, τ)の集合の中で与えられた評価関数を最大化し

することを問題とする。停止時間は停止規則と同等であるから，pair の政策 (π, τ) はまた，通常の意味の政策と停止則則との pair ともみなせる。

近来，ソ連でマルコフ過程論（特に強マルコフ過程）の発展に伴って，E. B. Dynkin および A. N. Shiryaev 等がいくつかの具体的な最適停止問題について，興味ある結果を与え，またアメリカでは古くは Y. S. Chow, H. Robbins，近年では L. E. Dubins, L. J. Savage 等により多くの具体的な最適停止問題が扱われて来た。

この報告は，定式化としては，前述のように通常の stochastic control problem に停止規則を入れた型でもあるし，また，別の面から見ると，上記の最適停止問題の一般化にさらに通常の control action を入れた型ともみなせる。

この報告は全体を通じて，上記の意味での pair の政策の class における最適政策（最適性の規準は多くいくつか与えられる）の存在定理と，最適政策の性質について論じたものである。その他，これらに付随した多くの定性的な興味ある結果が与えられている。最適解の構成に関しては，状態空間，行動空間をごく簡単なものに限定すれば容易であるが，この報告でとり上げたような一般的な距離空間

では，可成り困難な問題で，この報告でも explicit な型で述べられ所までは追究されていない。(存在定理の証明方法が有限 stage による政策改良の線に沿っているから，この意味では可成り constructive である)

以上のように，最適政策の構成法に関しては不十分であるが，存在定理 および最適政策の性質等に関しては，一応初期の目的を達したので，この報告で綜括する。

§2. 記号および定式化

2.1 一般的な定義

Borel set ; complete separable metric space の Borel subset
(X, Y 等で表わす)

Borel set X 上の probability measure ; X の Borel field 上の probability measure

P(X) ; X 上のすべての probability measure からなる空間

$q(y|x)$; x を与えたときの Y 上の conditional prob. measure で

(i) 各 $x \in X$ に対し $q(\cdot|x)$ は Y 上の prob. measure

(ii) 各 Borel subset $B \subset Y$ に対し $q(B|\cdot)$ は X 上の Baire 関数

$Q(Y|X)$; 定義空間 X, Y をきめて上のようなすべての conditional prob. measure からなる空間

$M(X)$; (i) §2 ～ §4 では X 上のすべての実数値 Baire 関数からなる空間

(ii) §5, §6 では X 上のすべての有界 Baire 関数からなる空間

$u, v \in M(X)$ に対して $u \geqq v$; $u(x) \geqq v(x)$ for all $x \in X$

$p \in P(X)$, $u \in M(X)$ に対して pu ; $pu = \int_X u(x)\, dp(x)$

$q \in Q(Y|X)$, $u \in M(XY)$ に対して qu ; $qu(x) = \int_Y u(x,y)\, dq(y|x)$

$p \in P(X)$ が degenerate ; $p\{x_0\} = 1$ for some $x_0 \in X$

$q \in Q(Y|X)$ が degenerate ; $q(\cdot|x)$ が各 x に対して degenerate

$q \in Q(Y|X)$ が degenerate であれば $q(\{f(x)\}|x) = 1$ for all $x \in X$ なる X から Y への Baire 関数 f が存在することになり, 従って $u \in M(XY)$ なる u に対しては

$$qu(x) = u(x, f(x)) \quad \text{for all } x \in X$$

となる。

2.2 決定過程に固有な定義

停止規則をともなう決定過程に対する最適化問題はつぎの 6 箇の要素 S, A, q, r, g, α で決定される.

S ; 状態空間 (state space) を表わし, ある空でない Borel set

A ; 行動空間 (action space) を表わし, ある空でない Borel set

$q=\{q_1, q_2, \cdots\}$; system の運動法則を表わし, 各 n につき

$$q_n \in Q(S|H_nA), \quad H_n = SASA\cdots S \quad (2n-1\text{ケ})$$

$r \in M(SAS)$; 利得関数 (reward function)

g ; 最終利得関数で S 上の有界 Baire 関数

α ; 割引率 (discount factor)

(i) §2～§4 では $0 \leqq \alpha \leqq 1$

(ii) §5, §6 では $0 \leqq \alpha < 1$

つぎに, policy, stopping time および stopped policy 等についての定義を述べる.

政策 (policy) $\pi = \{\pi_1, \pi_2, \cdots\}$; $\pi_n \in Q(A|H_n)$ $n=1,2,\cdots$

Markov policy ; $\pi=\{\pi_1, \pi_2, \cdots\}$ において各 π_n が $Q(A|S)$ の degenerate な要素

Markov policy を $\{f_1, f_2, \cdots\}$ (ただし f_n は S から A への Baire 関数) と書く.

定常政策 (stationary policy) ; $\pi=\{f_1, f_2, \cdots\}$ が Markov policy でかつ $f_n = f$ for all n

stationary policy を f^∞ と書く.

$\pi=\{\pi_1, \pi_2, \cdots\}$ に対して ${}^n\pi = \{\pi_{n+1}, \pi_{n+2}, \cdots\}$ と書く. 故に ${}^0\pi = \pi$

π を任意の policy とすれば, π と q とで つぎのような conditional probability measure e_π が定義される。

$$e_\pi = \pi_1 q_1 \pi_2 q_2 \cdots \in Q(ASAS\cdots \mid S)$$

$\mathcal{S}$; S の Borel field

$\mathcal{A}$; A の Borel field

標本空間 (sample space) ; $\Omega = SASA\cdots$

標本点 (sample point) 又は 道 (path) ; $\omega = (s_1, a_1, s_2, a_2, \cdots)$

$\mathcal{F}_n$; $\{\omega \mid s_1 \in E_1, a_1 \in F_1, s_2 \in E_2, a_2 \in F_2, \cdots, s_n \in E_n\}$

(ただし $E_i \in \mathcal{S}$, $F_i \in \mathcal{A}$, $i = 1, 2, \cdots, n$)

の型の ω - 集合のあらゆる和の族

($\mathcal{F}_n$ は 明らかに σ-field になる)

policy π に付随する停止時間 (stopping time)

$t = t(\omega)$;
- (i) 正整数値の確率変数
- (ii) $e_\pi[\{t(\omega) < \infty\}_{s_1} \mid s_1] = 1$ for all $s_1 \in S$
- (iii) $\{t(\omega) = n\} \in \mathcal{F}_n$ for each n

$\mathcal{C}(\pi)$; π に付随するすべての stopping time の集合

停止政策 (stopped-policy) (π, t) ;

π は任意の policy, $t \in \mathcal{C}(\pi)$ のとき (π, t) を stopped-policy といい, 簡単のために s-policy とかく

$\mathcal{C}^N(\pi)$; $\mathcal{C}(\pi)$ に属しかつ

$e_\pi[\{t(\omega) \leqq N\}_{s_1} \mid s_1] = 1$ for all $s_1 \in S$

をみたすすべての stopping time の集合

打切政策 (truncated-policy) ; s-policy (π, t) で あってかつ $t \in C^N(\pi)$ なる正整数 N が存在するもの を truncated-policy といい, 簡単のために t-policy とかく

$\mathcal{B}_n$; $\{\omega \mid s_1 \in E_1, s_2 \in E_2, \cdots, s_n \in E_n\}$ (ただし $E_i \in \mathcal{E}$ $i=1,2,\cdots,n$) の型の ω-集合のあらゆる和の族

($\mathcal{B}_n$ は明らかに σ-field)

$\mathcal{G}_n$; $\{\omega \mid s_n \in E\}$ (ただし $E \in \mathcal{E}$) の型のあらゆる ω-集合の族

$t \in C(\pi)$ が Markov stopping time ;

$$\begin{cases} \text{(i)} & \{t(\omega) = n\} \in \mathcal{B}_n \quad \text{for each } n \\ \text{(ii)} & \{t(\omega) = n\} = \{t(\omega) > n-1\} \cap \Delta_n \end{cases}$$

ただし $\Delta_n \in \mathcal{G}_n$, for each n

Markov s-policy (π, t) ; $\begin{cases} \text{(i)} & \pi \text{ が Markov policy} \\ \text{(ii)} & t \in C(\pi) \text{ かつ } t \text{ が Markov stopping time} \end{cases}$

stationary t-policy (π, t) ; (π, t) が Markov s-policy でかつ, π が stationary policy

$t \in C(\pi)$ が stationary stopping time ; t が Markov で
かつ Δ_n が n に無関数

つぎに評価関数の設定と，いくつかの最適性の規準を与える。

$$x_n = \sum_{k=1}^{n-1} \alpha^{k-1} r(s_k, a_k, s_{k+1}) + \alpha^{n-1} g(s_n), \qquad n=1,2,\cdots$$

これを簡単のために

$$x_n = \sum_{k=1}^{n-1} \alpha^{k-1} r_k + \alpha^{n-1} g_n, \qquad n=1,2,\cdots$$

とかくこともある。

E^π ; e_π に関する expectation を表わす積分作用素

s-policy (π, t) から得られる総期待利得を評価関数として採用する。すなわち

$$E^\pi(x_t) = E^\pi\left[\sum_{k=1}^{t-1} \alpha^{k-1} r_k + \alpha^{t-1} g_t\right]$$
$$= e_\pi\left[\sum_{k=1}^{t-1} \alpha^{k-1} r_k + \alpha^{t-1} g_t\right]$$

我々の目的は，上に定義した $E^\pi(x_t)$ を，何らかの意味で (π, t) について最大化することである。

Π^N ; N番 stage までに対するすべての policy $\{\pi_1, \pi_2, \cdots, \pi_N\}$ の集合

$$\Lambda^N \rightleftharpoons \{(\pi, t) \mid \pi \in \Pi^{N-1},\ t \in C^N(\pi)\}$$

Π ; 無限 stage にわたってのすべての policy の集合

$$\Lambda \rightleftharpoons \{(\pi, t) \mid \pi \in \Pi,\ t \in C(\pi)\}$$

$(p,\varepsilon)^N$-optimal ; $(\hat{\pi},\hat{\tau}) \in \Lambda^N$ かつ

$$P\{E^{\hat{\pi}}(x_{\hat{\tau}}) \geqq E^{\pi}(x_\tau) - \varepsilon\} = 1 \quad \text{for } {}^\forall(\pi,\tau) \in \Lambda^N$$

のとき $(\hat{\pi},\hat{\tau})$ を $(p,\varepsilon)^N$-optimal という

(p,ε,δ)-optimal ; $(\hat{\pi},\hat{\tau}) \in \Lambda$ かつ

$$P\{E^{\hat{\pi}}(x_{\hat{\tau}}) \geqq E^{\pi}(x_\tau) - \varepsilon\} \geqq 1-\delta \quad \text{for } {}^\forall(\pi,\tau) \in \Lambda$$

のとき $(\hat{\pi},\hat{\tau})$ を (p,ε,δ)-optimal という

(p,ε)-optimal ; $(\hat{\pi},\hat{\tau})$ が $(p,\varepsilon,0)$-optimal のとき，これを (p,ε)-optimal という

ε-optimal ; $(\hat{\pi},\hat{\tau}) \in \Lambda$ かつ

$$E^{\hat{\pi}}(x_{\hat{\tau}}) \geqq E^{\pi}(x_\tau) - \varepsilon \quad \text{for } {}^\forall(\pi,\tau) \in \Lambda$$

のとき $(\hat{\pi},\hat{\tau})$ を ε-optimal という

optimal ; $(\hat{\pi},\hat{\tau})$ が 0-optimal のとき これを optimal という

(p,ε)-dominate ; $P\{E^{\hat{\pi}}(x_{\hat{\tau}}) \geqq E^{\tilde{\pi}}(x_{\tilde{\tau}}) - \varepsilon\} = 1$ のとき $(\hat{\pi},\hat{\tau})$ は $(\tilde{\pi},\tilde{\tau})$ を (p,ε)-dominate するという

§3. (p,ε,δ)-optimal policy および (p,ε)-optimal policy の存在

先ず本論に入る前に $\{x_n\}$-process から $\{\beta_n^N(\pi)\}$-process

を構成し それにもとづく2つの proposition をあげる。

π を任意の policy として各 $N \geq 1$ に対し, $\{\beta_n^N(\pi)\}$-process を 次式で backward に定義する。

$$\begin{cases} \beta_n^N(\pi) = \max\left[x_n,\ E^{\pi_{n-1}}(\beta_{n+1}^N(\pi)) \right], & n=1,2,\cdots,N-1 \\ \beta_N^N(\pi) = x_N \end{cases} \tag{3.1}$$

ただし 各 x_n は E^π に関して可積分であるとし, $E^{\pi_{n-1}}$ は conditional prob. measure $\pi_n \xi_n \pi_{n+1} \xi_{n+1} \cdots$ に関する積分を表わす。したがって policy に関する仮定から, 各 N, n につき $\beta_n^N(\pi)$ は $\mathcal{F}_n$-可測である。

$\{\beta_n^N(\pi)\}$ を用いて τ_N を定義する。

$$\tau_N(\pi) = \text{the first } n \text{ such that } \beta_n^N(\pi) = x_n. \tag{3.2}$$

(3.1) により $\beta_N^N(\pi) = x_N$ だから $\tau_N(\pi) \in C^N(\pi)$.

今後簡単のために, 誤解の恐れのない限り $\tau_N(\pi)$ を τ_N で表わす。 τ_N の定義から明らかに

$$x_{\tau_N} = \beta_{\tau_N}^N(\pi).$$

Proposition 1 (3.1), (3.2) で定義された $\{\beta_n^N(\pi)\}$, τ_N に対して

(a) $E^\pi(x_{\tau_N}) = \sup_{t \in C^N(\pi)} E^\pi(x_t)$

(b) $E^\pi(\beta_{\tau_N}^N(\pi)) = \sup_{t \in C^N(\pi)} E^\pi(\beta_t^N(\pi))$

Proposition 2　(3.1), (3.2) で定義された $\{\beta_n^N(\pi)\}$, τ_N に対し

$$E^\pi(x_{\tau_N}) = E^\pi[\beta^N_{\tau_N}(\pi)] = \beta_1^N(\pi)$$

(注意) Proposition 1 および Proposition 2 は, この報告の本来の目的のためには全く準備的な事がらに過ぎないが, これ自体また別の意味で非常に興味ある結果である。これらの証明は かなりの補助定理を必要とするのでここでは省略するが, $\{\beta_n^N(\pi)\}$-process が super Martingale 性をもつことから Martingale system theory を用いることを付記しておく。

$X^N \equiv ASAS\cdots S$　(2N factors)

Σ^*; $P(X^N)$ の上の σ-field で 各 Borel subset $B \subset X^N$ に対し, 写像 $\nu(B)$; $P(X^N) \to R^1$ が Σ^*- 可測となる最小の σ- field

Σ^* は, weak-topology に関する $P(X^N)$ の Borel field と一致することが分る.

$\Gamma^N \equiv \{(s,\nu) \mid s \in S,\ \nu = e_\pi(s)$ for some $\pi \in \Pi^N\}$

Lemma 3.1　Γ^N は $S\cdot P(X^N)$ の Borel subset である。

次に $\{\beta_n^N(\pi)\}$-process に類似な方法で $\{\beta_n^N(\nu)\}$-process を構成する.

$\nu \in \mathcal{P}(X^N)$ を任意に与え、ν の分解を考える。

$$\nu = \nu_1 \nu_2 \cdots \nu_{2N},$$

ただし

$$\nu_1 \in \mathcal{P}(A)$$
$$\nu_2 \in Q(S \mid A)$$
$$\nu_{2n+1} \in Q(A \mid ASAS \cdots AS) \qquad (2n \text{ factors})$$
$$\nu_{2n} \in Q(S \mid ASAS \cdots A) \qquad (2n-1 \text{ factors}).$$

$\{\beta_n^N(\nu)\}$ を次式で backward に構成する。

$$\begin{cases} \beta_n^N(\nu) = \max\left[x_n,\ \nu_{2n-1}\nu_{2n}\beta_{n+1}^N(\nu)\right], & n=1,2,\cdots,N-1 \\ \beta_N^N(\nu) = x_N \end{cases} \tag{3.3}$$

ここで次の仮定をおく。

(A1) つぎの (a), (b), (c) のうち少くとも一つが成立すること。

(a) $-\infty < \beta_1^N(\nu) < \infty \quad \text{for } {}^\forall \nu \in \mathcal{P}(X^N)$

(b) $r \geqq 0, \quad q \geqq 0$

(c) $r \leqq 0, \quad q \leqq 0.$

(注意) (A1) の (b), (c) はそれぞれ、いわゆる positive case, negative case である。また (a) は §5, §6 で扱う discounted case では常に満足される。discounted case を今後、D-case と呼ぶ。

$$v^{N*} \rightleftharpoons \sup_{\pi \in \Pi^N} \beta_1^N(\pi).$$

一般に v^{N*} は可測になるとは限らないが、次の Lemma が成立つ。

Lemma 3.2　(A1) を仮定すると、v^{N*} は絶対可測である。

(略証)

$$v(\Delta,\nu) \equiv \beta_1^N(\nu)(\Delta)$$

$$B_\lambda \equiv \Gamma^N \cap \{(\Delta,\nu) \mid v(\Delta,\nu) > \lambda\}$$

とおくと Σ^* の作り方から $v(\Delta,\nu)$ が $SP(X^N)$ 上で (Δ,ν) に関して可測なることと、Lemma 3.1 により B_λ は $SP(X^N)$ の Borel subset になる。一方

$$v^{N*}(\Delta) = \sup_{\nu\in\Gamma_\Delta^N} v(\Delta,\nu) \qquad (\text{ただし } \Gamma_\Delta^N \text{ は } \Gamma^N \text{ の } \Delta\text{-section})$$

だから、$C_\lambda = \{\Delta \mid v^{N*}(\Delta) > \lambda\}$ は B_λ の S^Δ の projection である。従って C_λ は analytic set。故に絶対可測である。　(Q.E.D.)

Lemma 3.3　(A1) を仮定すると、任意の $p\in P(S)$、任意の $\varepsilon>0$ に対して

$$p\{\beta_1^N(\pi) \geqq v^{N*}-\varepsilon\} = 1$$

なるような $\pi\in\Pi^N$ が存在する。

(略証)

Lemma 3.2 により任意の $p\in P(S)$ に対して Borel set $N_1 \subset S$ と可測関数 v_0 が存在して

$$p(N_1)=0 \quad \text{かつ} \quad v_0(s)=v^{N*}(s) \quad \text{for } s\notin N_1.$$

つぎに

$$\Gamma_\varepsilon \equiv \Gamma^N \cap [\{(s,y) \mid s\notin N_1, v(s,y)>v_0(s)-\varepsilon\} \cup \{(s,y) \mid s\in N_1\}]$$

とおくと，Σ^* の定義と Lemma 3.1 により，Γ_ε は $SP(X^N)$ の Borel subset である。また

$$\Gamma_\varepsilon \text{ の } s\text{-section} = \phi \quad \text{for all } s\in S$$

は，Γ^N と v^{N*} の定義より明らか。故に G.W. Mackey の可測陰関数の定理（[5]の Theorem 6.3）により，Borel set $N_2\subset S$ と可測関数

$$y=\gamma(s) \; ; \; S\to P(X^N)$$

が存在して，$P(N_2)=0$ かつ $(s,\gamma(s))\in\Gamma_\varepsilon$ for $s\notin N_2$.

Σ^* の定義と γ の可測性から，各 $B\subset X^N$ に対して $\gamma(s)(B)$ は s に関して可測。故に $\mu(\cdot\mid s)=\gamma(s)(\cdot)$ とおけば，$\mu\in Q(X^N\mid S)$ となる。

$$\mu=\mu_1\mu_2\cdots\mu_{2N},$$
$$\mu_1\in Q(A\mid S)$$
$$\mu_{2n}\in Q(S\mid SA\cdots A) \qquad (2n \text{ factors})$$
$$\mu_{2n+1}\in Q(A\mid SA\cdots S) \qquad (2n+1 \text{ factors})$$

と分解出来る。かつ

$$\mu_{2n}(\cdot\mid s_1a_1\cdots s_na_n)=g_n(\cdot\mid s_1a_1\cdots s_na_n) \quad \text{for } s_1\notin N_2$$

となる。 $\hat{\pi} = \{\hat{\pi}_1, \hat{\pi}_2, \cdots, \hat{\pi}_N\}$ を

$$\hat{\pi}_n = \begin{cases} \mu_{2n-1} & \text{for } s_1 \notin N_2 \\ \pi_n' & \text{for } s_1 \in N_2 \end{cases} \qquad \text{ただし } \pi_n' \text{ は任意}$$

$$(n=1, 2, \cdots, N)$$

で定義すれば, $s_1 \notin N_1 \cup N_2$ ならば

$$v(s_1, \gamma(s_1)) = \beta_1^N(\hat{\pi})_{s_1} \geqq v_0(s_1) - \varepsilon = v^{N*}(s_1) - \varepsilon .$$

しかるに $p(N_1 \cup N_2) = 0$ だから

$$p\{\beta_1^N(\hat{\pi}) \geqq v^{N*} - \varepsilon\} = 1 \qquad \text{(Q.E.D.)}$$

(注意) Lemma 3.3 の証明で Mackey の陰関数の定理を用いたが, これが適用出来るためには $P(X^N)$ が位相空間として standard Borel space であれば十分である。(S はもともと standard Borel space である) $P(X^N)$ がこの§のはじめに導入した weak-topology に関して standard Borel space になることは K. R. Parthasarathy によって [6] の第2章に詳しく述べられている。

つぎの定理は有限 stage の model に対する最適政策の存在定理であるが, 実はこれがこの§で求める (p, ε, δ)-optimal s-policy の存在定理のための一つの重要な basis である。

Theorem 3.1　(A1) を仮定する。任意の $p \in P(S)$, 任意の $\varepsilon > 0$ に対して $(p, \varepsilon)^N$-optimal t-policy が存在する。

(略証)

Lemma 3.3 により任意の $p \in P(S)$, $\varepsilon > 0$ に対して

$$p\{\beta_1^N(\hat{\pi}) \geqq v^{N*} - \varepsilon\} = 1$$

なる $\hat{\pi} \in \Pi^{N-1}$ がある。故に

$$p\{\beta_1^N(\hat{\pi}) \geqq \beta_1^N(\pi) - \varepsilon\} = 1 \quad \text{for all } \pi \in \Pi^{N-1} \tag{3.4}$$

π を任意の policy として, $\tau_N(\hat{\pi})$, $\tau_N(\pi)$ をそれぞれ (3.2) で定義される $\hat{\pi}$ と π に対応する stopping time とする。

Proposition 2 により, (3.4) に代入して

$$p\{E^{\hat{\pi}}(x_{\tau_N(\hat{\pi})}) \geqq E^{\pi}(x_{\tau_N(\pi)}) - \varepsilon\} = 1 \quad \text{for } {}^{\forall}\pi \in \Pi^{N-1}$$

Proposition 1-(a) により

$$p\{E^{\hat{\pi}}(x_{\tau_N(\hat{\pi})}) \geqq E^{\pi}(x_t) - \varepsilon\} = 1 \quad \text{for } {}^{\forall}(\pi, t) \in \Lambda^N.$$

(Q.E.D.)

ここで仮定 (A1) に代えて, 仮定 (B1) をおく。

(B1)　つぎの (a), (b), (c) のうち少くとも一つが成立つこと。

(a)　(A1) の (a) が各 $N \geqq 1$ について成立つ

(b)　$\gamma \geqq 0$, $g \geqq 0$

(c)　$\gamma \leqq 0$, $g \leqq 0$

(注意)　(B1) の (a) は D-case では常にみたされる。

Lemma 3.4　(B1) を仮定する。p を $P(S)$ の任意の要素，ε も任意の正数とする。各 $N \geq 1$ に対して $(p,\varepsilon)^N$- optimal t-policy を $(\hat{\pi}^N, \hat{\tau}_N)$ とする。(その存在は Theorem 3.1 により保証される)。このとき一般性を失うことなく，p 確率 1 で $\{E^{\hat{\pi}^N}(x_{\hat{\tau}_N})\}$ は単調非減少であるとしてよい。

ここであらたに次の仮定をおく。

(A2)　$\sup_{(\pi,t)\in\Lambda} E^{\pi}(x_t) < \infty$

(A3)　$r_k = r'_k - r''_k$　　$(k=1,2,\cdots)$

ただし $r''_k \geq 0$ で r'_k, r''_k は共に SAS 上の実数値 Baire 関数.

(A4)　$E^{\pi}(x'_t) < \infty$ for every $(\pi,t)\in\Lambda$

ただし　$x'_t = \sum_{k=1}^{t-1} \alpha^{k-1} r'_k + \alpha^{t-1} g_t$.

(A5)　$\{(x'_n)^-,\ n\geq 1\}$ が各 e_π に関して一様可積分

ただし　$x'_n = \sum_{k=1}^{n-1} \alpha^{k-1} r'_k + \alpha^{n-1} g_n$.

(注意)　(A2) は D-case では常にみたされる。また (A3) において $r''_k \geq 0$ ということは r''_k が cost または cost を含む損失の部分を表わすと思えばよい。このとき r'_k には可測性の他に制限がないから，我々の

問題は cost だけを評価関数として cost を最小にする最適化問題をも含むことになる。この意味で仮定(A3)はごく一般的な仮定である。

以上の仮定のもとで次の Lemma を得る。

Lemma 3.5　(B1), (A2)～(A5) を仮定する。各 N に対する $(p,\varepsilon)^N$-optimal t-policy を $(\hat{\pi}^N, \hat{\tau}_N)$ で表わせば

$$P\left\{\lim_{N\to\infty} E^{\hat{\pi}^N}(x_{\hat{\tau}_N}) \geqq E^{\pi}(x_t) - \varepsilon\right\} = 1 \quad \text{for } \forall(\pi,t) \in \Lambda$$

が成立する。

(略証)

一般性を失うことなく, $E^{\pi}(x_t) \neq -\infty$ としてよい。

(A4) より　$-\infty < E^{\pi}(x_t) \leqq E^{\pi}(x'_t) < \infty$

$\therefore\ -\infty < E^{\pi}(x''_t) < \infty$　ただし $x''_t = \sum_{k=1}^{t-1} \alpha^{k-1} r''_k$　(3.5)

一方　$t \in C(\pi)$ に対して

$$t > n \Rightarrow x_n \geqq -(x'_n)^- - x''_t$$

故に

$$\int_{\{t>n\}} x_n^- \, de_\pi \leqq \int_{\{t>n\}} [(x'_n)^- + x''_t] \, de_\pi \qquad (3.6)$$

(3.5) (3.6) (A5) により

$$\lim_{n\to\infty} \int_{\{t>n\}} x_n^- \, de_\pi = 0 \qquad (3.7)$$

次に, $t_N = \min(t, N)$ とおけば, $(\pi, t_N) \in \Lambda^N$.

故に $(\hat{\pi}^N, \hat{\tau}_N)$ の定義により

$$\int_{\{t\leqq N\}} x_t\, de_\pi = E^\pi(x_{t_N}) - \int_{\{t>N\}} x_N\, de_\pi$$

$$\leqq E^{\hat{\pi}^N}(x_{\hat{\tau}_N}) + \varepsilon - \int_{\{t>N\}} x_N\, de_\pi \qquad \text{w.p.1}$$

$$\leqq E^{\hat{\pi}^N}(x_{\hat{\tau}_N}) + \varepsilon + \int_{\{t>N\}} \bar{x}_N\, de_\pi \qquad \text{w.p.1}$$

(3.8)

しかも (3.8) は各 $N\geqq 1$ について成立つ。(3.8) において $N\to+\infty$ とすれば (A2), Lemma 3.4, (3.7) により $\lim_{N\to\infty} E^{\hat{\pi}^N}(x_{\hat{\tau}_N})$ は存在して finite で, Lemma の求める結果が得られる。 (Q.E.D.)

この§の主目的である (p,ε,δ)-optimal s-policy の存在定理として つぎが与えられる。

Theorem 3.2 (B1), (A2)〜(A5) を仮定する。
各 $p\in P(S)$, 各 $\varepsilon>0$, 各 $\delta>0$ に対して (p,ε,δ)-optimal s-policy が存在し, かつそれは t-policy の class の中にある。

(略証)

Lemma 3.5 に Egoroff の定理を適用すれば 直ちに得られる。

Corollary 3.1　Theorem 3.2 と同じ仮定をおく。

(a) S が有限集合ならば，任意の $p \in P(S)$，任意の $\varepsilon > 0$ に対して (p, ε)-optimal t-policy が存在する。

(b) $p \in P(S)$ が finite support をもてば，任意の $\varepsilon > 0$ に対して (p, ε)-optimal t-policy が存在する。

(証明)

(a) または (b) の仮定があれば，p-確率 1 で $\{E^{\hat{\pi}^N}(x_{\hat{\tau}_N})\}$ が一様収束することから明らかである。

§4. Markov stopped-policy

この § を通じて system の運動法則は Markov type であるとする。即ち

$$q_n \in Q(S \mid SA) \quad \text{for each } n$$

とする。ただし $Q(S \mid SA)$ における $\mid$ の右側の SA は n 番 stage における状態空間と行動空間を表わすものとする。

§3 で (p, ε, δ)-optimal t-policy の存在を与えたが，この § では更にそれが Markov t-policy の class の中にあることを示す。そのために，次のように $\{\beta_n^N(\pi)\}$ を $\{v_n^N(\pi)\}$ に変換する。

$$\alpha^{n-1}v_n^N(\pi)=\beta_n^N(\pi)-\sum_{k=1}^{n-1}\alpha^{k-1}\gamma_k,\quad n=1,2,\cdots,N \qquad (4.1)$$

従って $\{\beta_n^N(\pi)\}$ の定義式 (3.1) は $\{v_n^N(\pi)\}$ を用いて次のように書き代えられる。

$$\begin{cases} v_n^N(\pi)=\max\left[g_n,\ \pi_n g_n\{\alpha v_{n+1}^N(\pi)+\gamma_n\}\right], & n=1,2,\cdots,N-1 \\ v_N^N(\pi)=g_N \end{cases} \qquad (4.2)$$

(注意) (4.2) をみると $\{\beta_n^N(\pi)\}$-process では, はっきりしなかった事が明らかになる。即ち (4.2) は policy π に従って行動してゆくとき, 最適な stopping rule によって得られる optimal return に関する dynamic programming form の再帰式であり, $v_n^N(\pi)$ は残りの stage 数が n であるときの optimal return を表わす。

次にあげる Lemma は §3 で引用した Mackey の陰関数の定理に匹敵する重要なもう一つの可測陰関数の定理である。

Lemma 4.1 (Blackwell and Ryll-Nardzewski [1])

X, Y を Borel set とする。任意の $q\in Q(Y|X)$, 任意の $w\in M(XY)$, 任意の $\varepsilon>0$ に対して

$$q[\{y;\ w(x,f(x))\geqq w(x,y)-\varepsilon\}\mid x]=1 \quad \text{for } \forall x\in X$$

なる Baire 関数 $f: X\to Y$ が存在する。

次の Lemma もまた 一つの存在関数の定理である。

Lemma 4.2　任意の π, $p \in P(S)$, $w \in M(SA)$, $\varepsilon > 0$ および各 n $(1 \leqq n \leqq N-1)$ に対して

$$p\pi_1 g_1 \cdots \pi_{n-1} g_{n-1} \{ f_n w \geqq \pi_n w - \varepsilon \} = 1$$

なる Baire 関数 $f_n ; S \to A$ が存在する。

(略証)

$$\mathcal{M} \rightleftharpoons p e_\pi \in P(SASA \cdots),$$

$\mathcal{M}$ の下で s_n が与えられたとの条件のもとでの a_n についての conditional prob. measure を π_n^* とすると, Lemma 4.1 により 任意の $w \in M(SA)$, $\varepsilon > 0$ および各 n に対して

$$\pi_n^* \{ w(s_n, f_n(s_n)) \geqq w(s_n, a_n) - \varepsilon \mid s_n \} = 1 \quad \text{for } \forall s_n \in S$$

なる Baire 関数 $f_n ; S \to A$ が存在する。故に

$$\begin{aligned} &\mathcal{M} \{ w(s_n, f_n(s_n)) \geqq w(s_n, a_n) - \varepsilon \} \\ &= p\pi_1 g_1 \cdots \pi_{n-1} g_{n-1} \pi_n^* \{ w(s_n, f_n(s_n)) \geqq w(s_n, a_n) - \varepsilon \mid s_n \} \\ &= p\pi_1 g_1 \cdots \pi_{n-1} g_{n-1} \cdot 1 = 1 \end{aligned}$$

(Q.E.D.)

Theorem 4.1　任意の $p \in P(S)$, $\varepsilon > 0$ および任意の t-policy (π, t) に対して, (π, t) を (p, ε)-dominate する Markov

t-policy が存在する。

(証明)

(π, t) を任意の t-policy とすると

$$(\pi, t) \in \Lambda^N \quad \text{for some } N.$$

(4.1) から

$$\beta_1^N(\pi) = v_1^N(\pi).$$

Proposition 1.2 により

$$E^\pi(x_t) \leqq E^\pi(x_{\tau_N(\pi)}) = \beta_1^N(\pi) = v_1^N(\pi) \tag{4.3}$$

(4.2) より

$$\begin{aligned} v_{N-1}^N(\pi_{N-1}) &= \max\left[g_{N-1},\ \pi_{N-1} g_{N-1} \{\alpha g_N + \gamma_{N-1}\} \right] \\ &= \max\left[g_{N-1},\ \pi_{N-1} w_{N-1} \right]. \end{aligned} \tag{4.4}$$

ここに

$$w_{N-1} = g_{N-1}\{\alpha g_N + \gamma_{N-1}\} \in M(SA).$$

$\zeta \rightleftharpoons \varepsilon / (1 + \alpha + \alpha^2 + \cdots + \alpha^{N-1})$ とおく。

Lemma 4.2 により Baire 関数 $f_{N-1} : S \to A$ がとれて

$$\pi_{N-1} w_{N-1} \leqq f_{N-1} w_{N-1} + \zeta \quad \text{with } p^{\pi_1 g_1 \cdots \pi_{N-2} g_{N-2}}\text{-prob. } 1$$

故に (4.4) から

$$\begin{aligned} v_{N-1}^N(\pi_{N-1}) &\leqq \max\left[g_{N-1},\ f_{N-1} w_{N-1} + \zeta \right] \\ &\qquad \text{with } p^{\pi_1 g_1 \cdots \pi_{N-2} g_{N-2}}\text{-prob. } 1 \\ &\leqq \max\left[g_{N-1}, f_{N-1} w_{N-1} \right] + \zeta \qquad 〃 \qquad 〃 \\ &= v_{N-1}^N(f_{N-1}) + \zeta \qquad 〃 \qquad 〃 \end{aligned}$$

$\therefore\ \pi_{N-2} g_{N-2}\{\alpha v_{N-1}^N(\pi_{N-1})+\gamma_{N-2}\}$

$\leqq \pi_{N-2} w_{N-2} + \alpha\varepsilon$ with $p\pi_1 g_1 \cdots \pi_{N-3} g_{N-3}$-prob. 1

ただし

$$w_{N-2} = g_{N-2}\{\alpha v_{N-1}^N(f_{N-1})+\gamma_{N-2}\} \in M(SA).$$

再び Lemma 4.2 により Baire 関数 $f_{N-2}; S\to A$ がとれて,

$$v_{N-2}^N(\pi_{N-2},\pi_{N-1}) = \max\left[g_{N-2}, \pi_{N-2} g_{N-2}\{\alpha v_{N-1}^N(\pi_{N-1})+\gamma_{N-2}\}\right]$$

$\leqq \max[g_{N-2}, f_{N-2} w_{N-2} + \varepsilon(1+\alpha)]$ with $p\pi_1 g_1 \cdots \pi_{N-3} g_{N-3}$-prob. 1

$\leqq v_{N-2}^N(f_{N-2}, f_{N-1}) + \varepsilon(1+\alpha)$ 〃 〃

以下同様に Baire 関数 $f_{N-3}, \cdots, f_1$ をえらんで

$$v_1^N(\pi) \leqq v_1^N(f_1, f_2, \cdots, f_{N-1}) + \varepsilon \quad \text{with } p\text{-prob. 1} \qquad (4.5)$$

$\pi^* \rightleftharpoons \{f_1, f_2, \cdots, f_{N-1}\}$

$\tau^* \rightleftharpoons$ the first n such that $v_m^N(\pi^*) = g_n$

とおくと

$\tau^* = \tau_N(\pi^*)$. (ただし $\tau_N(\pi^*)$ は (3.2) で定義される)

かつ, (π^*, τ^*) は Markov t-policy である。

(4.3) により (4.5) を書き直して

$$p\{E^{\pi^*}(x_{\tau^*}) \geqq v_1^N(\pi) - \varepsilon\} = 1$$

再び (4.3) により

$$p\{E^{\pi^*}(x_{\tau^*}) \geqq E^{\pi}(x_t) - \varepsilon\} = 1.$$

(Q.E.D.)

Theorem 4.2　Theorem 3.2 と同じ仮定をおく。
任意の $p \in P(S)$, 任意の $\varepsilon > 0$, 任意の $\delta > 0$ に対して,
(p, ε, δ)- optimal Markov π-policy が存在する。

(略証)

Theorem 3.2 と Theorem 4.1 より 明らか。

Corollary 4.1　Theorem 3.2 と同じ仮定をおく。

(a) S が有限集合であれば, 任意の $p \in P(S)$, $\varepsilon > 0$ に対して, (p, ε)-optimal Markov π-policy が存在する。

(b) $p \in P(S)$ が finite support をもてば, 任意の $\varepsilon > 0$ に対して, (p, ε)-optimal Markov π-policy が存在する。

(注意)　Theorem 4.1 の証明中で, $v_n^N(\pi)$ と書くべき所を $v_{N-1}^N(\pi_{N-1})$, $\pi_{N-2}^N(\pi_{N-2}, \pi_{N-1})$ 等と書いたが, 実は $v_n^N(\pi)$ は $\pi = \{\pi_1, \pi_2, \cdots\}$ の成分の中, $(\pi_n, \pi_{n+1}, \cdots, \pi_{N-1})$ のみに depend するからである。

§5　Stationary stopped-policy

この§では. D-case に対して, stationary s-policy の存在と, それに関連して再適方程式 (optimality equation)

等について論じる。

この §では つぎの仮定をおく。

(i) $0 \leqq \alpha < 1$

(ii) $M(X)$ は X 上のすべての有界 Baire 関数からなる空間を表わす。 $M(X)$ におけるノルムを $\|u\| = \sup_{x \in X} |u(x)|$ で定義する。

(iii) $g_1 = g_2 = g_3 = \cdots = g \in Q(S|SA)$.

次に degenerate $f \in Q(A|S)$ に対して operator T_f; $M(S) \to M(S)$ も次式で定義する。更に A_f を定義する。

$$T_f u = f g(r + \alpha u),$$

$$A_f u = \max[g, T_f u].$$

Markov policy $\pi^N = \{f_1, f_2, \cdots f_N\}$ $(N < \infty)$ に対して operator U_{π^N}, L_{π^N} ; $M(S) \to M(S)$ を次式で定義する。

$$U_{\pi^N} u = \max_{1 \leqq n \leqq N} T_{f_n} u,$$

$$L_{\pi^N} u = \max[g, U_{\pi^N} u].$$

これらの operator に関して直ちに 次の2つの Lemma を得る。

Lemma 5.1

(a) T_f, A_f, U_{π^N}, L_{π^N} は monotone operator である。

(b) $u \in M(S)$, 定数 c に対して

$$T_f(u+c) = T_f u + \alpha c, \quad U_{\pi^N}(u+c) = U_{\pi^N} u + \alpha c.$$

(c) $u \in M(S)$, 定数 $c > 0$ に対して

$$A_f(u+c) \leqq A_f u + \alpha c, \quad L_{\pi^N}(u+c) \leqq L_{\pi^N} u + \alpha c.$$

(d) $u \in M(S)$, 定数 $c < 0$ に対して

$$A_f(u+c) \geqq A_f u + \alpha c, \quad L_{\pi^N}(u+c) \geqq L_{\pi^N} u + \alpha c.$$

Lemma 5.2 (a) T_f, A_f, U_{π^N}, L_{π^N} は $M(S)$ 上の contraction mapping で その contraction coefficient は α である。

(b) $u, v \in M(S)$, Markov policy $\pi = \{f_1, f_2, \cdots\}$ に対して

$$\lim_{n\to\infty} \| A_{f_1} A_{f_2} \cdots A_{f_n} u - A_{f_1} A_{f_2} \cdots A_{f_n} v \| = 0.$$

次に π^N-generated の定義を与える。

Markov policy $\pi^N = \{f_1, f_2, \cdots, f_N\}$ に対して

$f(S \to A)$ が π^N-generated ; S の Borel 分割 $S_1, S_2, \cdots, S_N$ が有って $f = f_n$ on S_n

Markov policy $\hat{\pi}^N = \{g_1, g_2, \cdots, g_N\}$ が π^N-generated ; 各 g_n が π^N-generated

$F(\pi^N)$; すべての π^N-generated function の集合

$G^N(\pi^N)$; Π^N に属するすべての π^N-generated Markov policy の集合

Lemma 5.3

(a) $T_f u \leqq U_{\pi^N} u$ for any $u \in M(S)$, for any $f \in F(\pi^N)$

(b) $A_f u \leqq L_{\pi^N} u$ for any $u \in M(S)$, for any $f \in F(\pi^N)$

Lemma 5.4 任意の $u \in M(S)$ に対して次のような $f \in F(\pi^N)$ が存る：

$$T_f u = U_{\pi^N} u \quad \text{and} \quad A_f u = L_{\pi^N} u .$$

Lemma 5.5 u_N^* を L_{π^N} の fixed point とする。

(a) 任意の $\varepsilon > 0$ に対して N を十分大きくとれば

$$\beta_1^{N+1}(\hat{\pi}^N) - \varepsilon \leqq u_N^* \quad \text{for } \hat{\pi}^N \in G^N(\pi^N).$$

(b) 任意の $\varepsilon > 0$ に対して N を十分大きくとれば

$$\beta_1^{N+1}(f^{(N)}) \geqq u_N^* - \varepsilon$$

なる $f \in F(\pi^N)$ をえらぶことが出来る。ただし $f^{(N)} = \{f, f, \cdots, f\}$ (N factors)

(略証)

(a) $\hat{\pi}^N = \{\hat{f}_1, \hat{f}_2, \cdots, \hat{f}_N\} \in G^N(\pi^N)$ とする。

Lemma 5.3 (b) より

$$A_{\hat{f}_i} u \leqq L_{\pi^N} u \quad \text{for } 1 \leqq i \leqq N, \text{ for } u \in M(S)$$

$$\therefore A_{\hat{f}_N} u_N^* \leqq L_{\pi^N} u_N^* = u_N^*$$

$$\therefore A_{\hat{f}_1} A_{\hat{f}_2} \cdots A_{\hat{f}_N} u_N^* \leqq u_N^* \qquad (5.1)$$

Lemma 5.2 (b) と (5.1) により. N を十分大きくとれば"

$$u_N^* \geqq A_{\hat{f}_1} A_{\hat{f}_2} \cdots A_{\hat{f}_N} g - \varepsilon$$
$$= v_1^{N+1}(\hat{\pi}^N) - \varepsilon = \beta_1^{N+1}(\hat{\pi}^N) - \varepsilon.$$

(b) Lemma 5.4 (b) より $f \in F(\pi^N)$ があって

$$A_f u_N^* = L_{\pi^N} u_N^* = u_N^*.$$

$\therefore (A_f)^N u_N^* = u_N^*$

Lemma 5.2 (b) により N を十分大きくとれば"

$$(A_f)^N g + \varepsilon \geqq (A_f)^N u_N^* = u_N^* \qquad (5.2)$$

しかるに $(A_f)^N g = v_1^{N+1}(f^{(N)}) = \beta_1^{N+1}(f^{(N)})$ だ"から (5.2) より,

$$\beta_1^{N+1}(f^{(N)}) \geqq u_N^* - \varepsilon.$$

(Q.E.D.)

Theorem 5.1 (A3), (A4), (A5) を仮定する。このとき任意の $p \in P(S)$, 任意の $\varepsilon > 0$, 任意の $\delta > 0$ に対して (p, ε, δ)-optimal stationary s-policy が t-policy として存在する。

(略証)

$(\pi^{N_1}, \tau_{N_1}) \in \Lambda^{N_1}$ を $(p, \frac{\varepsilon}{3}, \delta)$-optimal Markov t-policy とする。(存在は Theorem 4.2)

§3 における議論から (π^{N_1}, τ_{N_1}) は $(p, \frac{\varepsilon}{3})^{N_1}$-optimal として得られる。

一方, 各 N に対する $(p, \frac{\varepsilon}{3})^N$-optimal t-policy を

(π^N, τ_N) とすれば Lemma 3.5 より

$$P\left\{E^{\pi^N}(x_{\tau_N}) \geqq E^{\pi}(x_t) - \frac{\varepsilon}{3}\right\} \geqq 1-\delta \quad \text{for } N \geqq N_1, \quad \text{for } (\pi, t) \in \Lambda \qquad (5.3)$$

L_{π^N} の fixed point を u^*_{N-1} とすれば Lemma 5.5 (b) より N_2 を十分大きくとれば, 各 $N \geqq N_2$ に対して

$$\exists f \in F(\pi^N)\ ;\ \beta_1^N(f^{(N-1)}) \geqq u^*_{N-1} - \frac{\varepsilon}{3} \qquad (5.4)$$

Lemma 5.5 (a) より N_3 を十分大きくとれば, 各 $N \geqq N_3$ に対して

$$u^*_{N-1} \geqq \beta_1^N(\pi^N) - \frac{\varepsilon}{3} \qquad (5.5)$$

$N_0 = \max[N_1, N_2, N_3]$ とおけば (5.4) (5.5) より

$$\beta_1^{N_0}(f^{(N_0-1)}) \geqq u^*_{N_0-1} - \frac{\varepsilon}{3} \geqq \beta_1^{N_0}(\pi^{N_0}) - \frac{2}{3}\varepsilon \qquad (5.6)$$

ここで

$$\tilde{\tau}_{N_0} = \text{the first } n \text{ such that } \beta_n^{N_0}(f^{(N_0-1)}) = x_n$$

とおけば

$$\tilde{\tau}_{N_0} \in C^{N_0}(f^{(N_0-1)}) \text{ かつ } \beta_1^{N_0}(f^{(N_0-1)}) = E^{f^{(N_0-1)}}(x_{\tilde{\tau}_{N_0}})$$

故に (5.6) より

$$E^{f^{(N_0-1)}}(x_{\tilde{\tau}_{N_0}}) \geqq E^{\pi^{N_0}}(x_{\tau_{N_0}}) - \frac{2}{3}\varepsilon \qquad (5.7)$$

(5.3) (5.7) より

$$P\left\{E^{f^{(N_0-1)}}(x_{\tilde{\tau}_{N_0}}) \geqq E^{\pi}(x_t) - \varepsilon\right\} \geqq 1-\delta \quad \text{for } \forall (\pi, t) \in \Lambda$$

即ち $(f^{(N_0-1)}, \tilde{\tau}_{N_0})$ は (p, ε, δ)-optimal stationary t-policy である。 (Q.E.D.)

Corollary 5.1　Theorem 5.1 と同じ仮定をおく。

(a) S が有限集合なら，任意の $p \in P(S)$, $\varepsilon > 0$ に対して (p, ε)-optimal stationary t-policy が存在する。

(b) $p \in P(S)$ が finite support をもてば，任意の $\varepsilon > 0$ に対して，(p, ε)-optimal stationary t-policy が存在する。

次に Markov policy $\pi = \{f_1, f_2, \cdots\}$ に対して operator U_π, L_π を定義する。即ち

$$U_\pi u = \sup_n T_{f_n} u, \qquad u \in M(S)$$
$$L_\pi u = \max[g, U_\pi u].$$

このとき，U_π, L_π は monotone operator で，かつ $M(S)$ 上の contraction mapping になる。

Markov policy $\pi = \{f_1, f_2, \cdots\}$ に対して

$f : S \to A$ が π-generated ; S の Borel partition $S_1, S_2, \cdots$ が有って $f = f_n$ on S_n.

Markov policy $\hat{\pi} = \{g_1, g_2, \cdots\}$ が π-generated ; 各 g_n が π-generated

$F(\pi)$; すべての π-generated function の集合

$G(\pi)$; π に係るすべての π-generated M-policy の集合

$\Lambda_\pi \equiv \{(\hat{\pi}, \hat{t}) \mid \hat{\pi} \in G(\pi),\ \hat{t} \in C(\hat{\pi})\}$

Optimality equation

A_a を $f \equiv a$ なる f に対応する operator とする。このとき $u \in M(S)$ に対して

$$u = \sup_{a \in A} A_a u \qquad (5.8)$$

が成立つとき，u は optimality equation をみたすといい，(5.8) を optimality equation という。

Optimal return

$u \in M(S)$ が

(i) $u \geqq E^{\hat{\pi}}(x_{\hat{\xi}})$ for $\forall (\hat{\pi}, \hat{\xi}) \in \Lambda_\pi$

(ii) $\forall \varepsilon > 0$, $\exists (\hat{\pi}, \hat{\xi}) \in \Lambda_\pi$; $E^{\hat{\pi}}(x_{\hat{\xi}}) \geqq u - \varepsilon$

をみたすとき，u を optimal return in Λ_π という。

$u \in M(S)$ が

(iii) $u \geqq E^{\hat{\pi}}(x_{\hat{\xi}})$ for $\forall (\hat{\pi}, \hat{\xi}) \in \Lambda$

(iv) $\forall \varepsilon > 0$, $\exists (\hat{\pi}, \hat{\xi}) \in \Lambda$; $E^{\hat{\pi}}(x_{\hat{\xi}}) \geqq u - \varepsilon$

をみたすとき，u を optimal return という。

Theorem 5.2 (A3), (A4), (A5) をおく。

(a) π を任意の policy，L_π の fixed point を u^* とすれば u^* は optimal return in Λ_π である。

(b) $\varepsilon \geqq 0$ とする。もし ε-optimal s-policy があれば

各 $\varepsilon'>0$ に対して $(\varepsilon/(1-\alpha)+\varepsilon')$-optimal stationary s-policy が存在する。

(c) $u \in M(S)$ が $A_a u \leqq u$ for all $a \in A$ をみたせば $E^{\pi}(x_t) \leqq u$ for $(\pi, t) \in \Lambda$ が成立する。

(d) 各 $\varepsilon>0$ に対して ε-optimal s-policy があれば optimal return は optimality equation をみたす。

(証明は非常に長いので予定く)

π が semi-Markov ; $\pi=\{\pi_1, \pi_2, \cdots\}$ において $\pi_n \in \Omega(A|SS)$

かつ π_n は degenerate

semi-Markov policy を用いて, 次の定理が導かれる。

Theorem 5.3 (A3) (A4) (A5) をおく.

s-policy (π^*, t^*) が optimal

$\Longleftrightarrow$ $E^{\pi^*}(x_{t^*})$ が optimality equation をみたす

Theorem 5.4 optimal equation は高々一つの bounded solution をもつ。

§6 Additional results.

特に, action space が countable set または finite set のときを扱う。

Theorem 6.1 (A3) (A4) (A5) をおく。

(a) A が countable set であれば，任意の $\varepsilon>0$ にたいして ε-optimal stationary s-policy が存在する。

(b) (A3) (A5) の他に（(A4) は不要），(i) $\lim_{n\to\infty} x''_n \equiv \lim_{n\to\infty} \sum_{k=1}^{n} \alpha^{k-1} y''_k = +\infty$，(ii) $\{x^-_n\}$ が各 e_π について一様可積分，(iii) $\sup_N E^{f^\infty}(x'_{\tau_N(f^\infty)}) < \infty$ for $\forall f^\infty$ の仮定をおく。このとき A が finite set なら stationary stopping time をもつような optimal stationary s-policy が存在する。

[付記]

この報告は [2] からの抜粋で，ここに紹介しなかったかなりの部分，および証明を除いた箇所については [2] を見て頂きたい。

直お §6 は optimal 又は ε-optimal s-policy の存在定理であるが，action space が countable でなくても，一般に compact であれば，他に transition probability q に弱連続の仮定等を加えれば Theorem 6.1 の結果と同様の結果が得られることが，最近 S. Iwamoto によって証明された。([3])

REFERENCES

[1] D. Blackwell and C. Ryll-Nardzewski ; Non-existence of everywhere proper conditional distributions. Ann. Math. Statist. 34 (1963), 223–225.

[2] N. Furukawa and S. Iwamoto ; Stopped decision processes on complete separable metric spaces. (to appear)

[3] S. Iwamoto; Stopped decision processes on compact metric spaces. (to appear)

[4] H.J. Kushner ; Computational procedures for optimal stopping problems for Markov chains. Jour. Math. Anal. Appl. 25 (1969), 607–615.

[5] G.W. Mackey ; Borel structure in groups and their duals. Trans. Amer. Math. Soc. 85 (1957), 134–165.

[6] K.R. Parthasarathy; Probability measures on metric spaces. (1967), Academic Press.